

# The case–combined-control design was efficient in detecting gene–environment interactions

N. Andrieu<sup>a,b,\*</sup>, A.M. Goldstein<sup>c</sup>

<sup>a</sup>Inserm EMI00-06, Tour Evry 2, 523 Place des Terrasses de l'Agora, 91034 Evry Cedex, France

<sup>b</sup>Service de Biostatistiques, Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 5, France

<sup>c</sup>Genetic Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20892, USA

Accepted 11 November 2003

## Abstract

**Objective:** The interest in studying gene–environment ( $G \times E$ ) interaction is increasing for complex diseases. A design combining both related and unrelated controls (e.g., population-based and siblings) is proposed to increase the power to detect  $G \times E$  interaction.

**Study Design and Setting:** We used simulations to assess the efficiency of the case–combined–control design relative to a classical case–control study under a variety of assumptions.

**Results:** The case–combined–control design appears more efficient and feasible than a classical case–control study for detecting interaction involving rare exposures and/or genetic factors. The number of available sibling controls per case and the frequencies of the risk factors are the most important parameters for determining relative efficiency. Relative efficiencies decrease as the frequency of the gene ( $G$ ) increases. A positive correlation in exposure ( $E$ ) between siblings decreases relative efficiency.

**Conclusions:** Although the case–combined–control design may not be efficient for common genes with moderate effects, it appears to be a useful alternative in certain situations where classical approaches remain unrealistic. © 2004 Elsevier Inc. All rights reserved.

**Keywords:** Case–control study;  $G \times E$  interaction; Population-based controls; Sibling controls; Study design

## 1. Introduction

The interest in studying gene–environment ( $G \times E$ ) interaction is increasing for complex diseases, particularly as molecular genetic technology improves. Moreover, detecting genetically homogeneous subgroups who are “susceptible” to specific environmental agents such as drugs, infectious agents, or life habits (e.g., smoking, diet) may have a public health impact through prevention or early screening actions [1]. However, few study designs have been fully evaluated for efficiency and power to detect gene–environment interaction [2–4]. Among the designs (e.g., the case–control or cohort design) that have already been evaluated for assessing  $G \times E$  interaction, most appear inefficient for detecting interaction involving rare environmental exposure(s) and/or genetic factors, particularly for moderate values of the  $G \times E$  interaction effect. Among the approaches that allow for simultaneously assessing the main effects and the interaction effect, usually either population-based or related controls serve as the referent group.

Designs using unrelated controls (e.g., population-based), like the classical case–control design or the cohort design, have the major disadvantage of very low power when the  $G \times E$  interaction involves rare exposure(s) and/or genetic factors. Indeed, detecting interaction has been shown to require at least four times as many subjects as detecting a main effect [5], and required sample sizes are often unattainable. Meta-studies and meta-analyses may provide approaches to rapidly increase sample sizes. However, although this methodology has been applied to traditional epidemiology studies, specific methods to deal with complex genetic issues have still to be fully developed [6]. Alternative approaches for detecting interaction including multistage designs [7,8] and flexible matching strategies [9,10] have been proposed to study rare factor(s). The goal of these approaches is to increase the frequency of the rare factor through oversampling [8]. One of these alternative designs, the counter-matching design appears to be more appropriate than most traditional methods for the study of  $G \times E$  interaction involving a rare factor [11]. However, its feasibility remains unrealistic for an event with a frequency of less than 0.01 and a moderate interaction effect (i.e.,  $<5$ ). Recently, flexible matching strategies with varying proportions of a matching factor among selected controls were

\* Corresponding author. Tel.: (33) 1 55.43.14.63; fax: (33) 1 55.43.14.69.

E-mail address: nadine.andrieu@curie.net (N. Andrieu).

shown to increase the power and efficiency of case–control studies to detect and estimate  $G \times E$  interactions compared to traditional frequency matching [9,10]. However, genetic and environmental factors with frequencies of less than 5 and 10%, respectively, could not be evaluated because of the small simulated sample sizes.

The use of unrelated controls to assess the genetic and  $G \times E$  effects has been questioned because of the potential problem of population stratification [12]. This potential bias from stratification was thus the motivation for some authors to propose related controls as a more appropriate control group for evaluating genetic factors. Witte et al. [13] showed that using population-based controls was more efficient in detecting a genetic main effect than using either cousins or siblings, with sibling controls being the least efficient. In contrast, sibling controls were the most efficient for detecting a  $G \times E$  interaction effect. This gain in efficiency decreased as the frequency of the genetic factor increased [13]. Additional studies [14] showed that the case–sibling design was most efficient when studying a dominant gene, whereas a case–parent design was preferred for a recessive gene. Other authors have been motivated to use related controls to get information about unknown within-family correlated factors associated with both the studied disease and exposure [15,16].

We propose a design using both related and unrelated controls (simultaneously), named the case–combined–control design, to increase the power to detect  $G \times E$  interaction without increasing dramatically the number of required study subjects. In addition, this design permits estimation of both the  $G \times E$  interaction and main effects. However, the purpose of this article is to examine the power in detecting interaction (which is limited to departure from multiplicative joint effects of  $G$  and  $E$  in this exercise); thus, power to detect main effects is not presented.

## 2. Methods

The population for the case–combined–control design consists of cases and two types of controls, unrelated controls and sibling controls. The unrelated controls may be population-based or recruited from electoral rolls, random digit dialing, neighborhoods, etc. Advantages and disadvantages of these different sources of controls have been discussed elsewhere [17–19].

Limited examinations have suggested that approximately 50% of cases may have appropriate sibling controls ([20,21], unpublished data). For purposes of presentation, we use this observation as the average number of available controls per case (defined as  $F$ ). Thus, in most of our evaluations of the proposed design,  $F = 0.5$ , meaning that approximately half of the cases have a sibling control. We note that the time period during which a sibling is eligible to be a control should be the time period in which that sibling is also eligible to be a case, should disease occur. If matching on age also

is required, it would be preferable to match sibling controls from specific age categories [22,23].

For purposes of presentation, we make several assumptions about the study population. We assume homogeneity between the odds ratios of the variables involved in the  $G \times E$  interactions using either of the two types of controls. We further assume that there is no population stratification bias. Finally, we assume that there is no difference in the distribution of variables of interest between cases who have sibling controls vs. those cases without such sibling controls, and that there is exchangeability of covariates of interest in cases and sibling controls, that is, that the covariate distribution does not depend on calendar time or birth order [24]. Then, combining the population-based and sibling control groups leads to an increase in the frequency of  $G$ , and in the frequency of  $E$  when there is a correlation between the case–sibling controls' exposures. Given the above assumptions, the proposed analysis for the case–combined–control design is a matched analysis. As such, each matched set would include a case, an unrelated control, and an unaffected sibling of the case (for the approximately 50% of cases).

To assess the proposed design, we compared the case–combined–control design to a classical case–unrelated–control study using simulations. The parameter of interest is the interaction odds ratio ( $R_I$ ) defined on a multiplicative scale. We define the parameters for modeling an interaction between a genetic factor  $G$  and an environmental exposure  $E$  below. Let

$P_E = P(E)$  = prevalence of the environmental

factor  $E$  in the population

$P_G = P(G)$  = prevalence of the genetic

factor  $G$  in the population

We further define the genetic factor  $G$  as follows. Let the alleles at the locus be classified as  $A$  (mutant) or  $a$  (wild), with population frequency  $p$  of the  $A$  allele and population frequency  $q$  for the  $a$  allele, where  $p + q = 1$ . We define dominant and recessive inheritance models. For the dominant model,  $AA$  and  $Aa$  represent subjects with  $G$ , and  $aa$  represents subjects with  $G$  under the recessive model. Thus,  $P_G = p^2 + 2pq$  for the dominant model, and  $P_G = p^2$  for the recessive model.

Let  $G$  and  $E$  be independent events. Further, we let the penetrance for  $G$  be incorporated directly into the risks below. Then,

$a = P(D|G^+, E^+)$  = risk of disease given

a person has  $G$  ( $G^+$ ) and  $E$  ( $E^+$ )

$b = P(D|G^+, E^-)$  = risk of disease given

a person has  $G^+$  and  $E^-$

$c = P(D|G^-, E^+) =$  risk of disease given

a person has  $G^-$  and  $E^+$

$d = P(D|G^-, E^-) =$  risk of disease given

a person does not have  $G$  or  $E$  ( $G^-, E^-$ )

Finally,

$R_E =$  odds ratio between  $E$  and disease

(among those not having  $G$ )

$R_G =$  odds ratio between  $G$  and disease

(among those not exposed to  $E$ )

$R_I =$  interaction effect, defined on a multiplicative scale.

Table 1 shows the subgroups of cases and unrelated controls at different risks for disease when there is a  $G \times E$  interaction under dominant and recessive genetic models derived from Smith and Day [5]. Conditional on the case genotypes, the genotype distributions of the case siblings are calculated and shown in Table 2. When there is a correlation in  $E$  between siblings, the probability that a case's sibling is exposed to  $E$  is defined as in Goldstein et al. [25]. Details for the calculations are presented in the Appendix.

### 3. Simulation studies

We calculated the expected distributions of the environmental exposure and genetic susceptibility in cases, unrelated, and related controls according to the parameters defined above and in Tables 1 and 2. Random numbers were generated to determine the number of controls for each case, that is, one unrelated control for all cases and one related control for approximately 50% of the cases for the case–combined–control study; one unrelated control for all cases and a second unrelated control for approximately 50% of the cases for the classical case–control study when  $F = 0.5$ . When  $F = 1.5$ , as another example, there are one unrelated control and one related control for all cases and

a second related control for approximately 50% of the cases for the case–combined–control study. For the classical case–control study, there are two unrelated controls for all cases and a third unrelated control for approximately 50% of the cases. When  $E$  and  $G$  were relatively common (e.g., both  $>0.05$ ), we simulated 2,500 data sets with 1,000 cases:1,000 unrelated controls: approximately 500 sibling controls when  $F = 0.5$ . When  $E$  and  $G$  were relatively rare (e.g., either  $<0.05$ ) (or very rare; e.g., both  $\leq 0.01$ ), we simulated 1,000 data sets with 5,000 (or 10,000) cases: 5,000 (or 10,000) unrelated controls: approximately 2,500 (or 5,000) sibling controls. In addition, a second set of 1,500 or 7,500 (or 15,000) unrelated controls was matched to the cases to conduct a classical case–unrelated–control study. The related controls are matched to the cases on family because they are siblings and the unrelated control are matched to the cases only on the number of strata. All subjects were simulated using random numbers generated by the SAS function RANUNI (SAS, version 8, Cary, NC) to assign each of the cases and controls to the different possible  $E$  and  $G$  categories.

Each simulated case–control study was analyzed with conditional logistic regression (CLR) using the program STATA [26] with a binary variable for  $E$  and a binary variable for  $G$  (based on the genotypes and inheritance model). When simulations of the combined design were performed under the null hypothesis ( $H_0$ ), the empirical Type I error rate for  $R_I$  approximated 0.05 suggesting that CLR is a reliable method for this design. To insure that CLR is an adequate method for analyzing the combined design, we also used an analytical approach to calculate the conditional Maximum Likelihood Estimate (MLE) of the ORs under  $H_0$  using a robust Mantel-Haenszel formula [27]. For simplicity, we generalized the correlation in  $G$  between relatives [25]. For any frequency of or correlation in  $G$ , the MLE for  $G$  and  $G \times E$  under  $H_0$  equaled 1, showing that the CLR appeared unbiased (data not shown).

We defined the relative efficiency ( $RE$ ) of the case–combined–control study compared to a classical case–control study, as the ratio of the variances of  $\beta_I$ , that is, the variance of  $\beta_I$  of the classical case–control study divided by the variance of  $\beta_I$  of the case–combined–control study. We used the same case:control ratio, that is, number of cases/number of

Table 1

Subgroups of the population at different risk of disease when there is a  $G \times E$  interaction according to Smith and Day [5]

Exposure	Proportion of controls	Dominant model		Recessive model	
		Relative disease risk	Proportion of cases	Relative disease risk	Proportion of cases
$E^+ [AA]$	$P_E p^2$	$R_E R_G R_I$	$(P_E p^2 R_E R_G R_I) / \Sigma^a$	$R_E R_G R_I$	$(P_E p^2 R_E R_G R_I) / \Sigma^a$
$E^+ [Aa]$	$P_E 2p(1-p)$	$R_E R_G R_I$	$(P_E 2p(1-p) R_E R_G R_I) / \Sigma$	$R_E$	$(P_E 2p(1-p) R_E) / \Sigma$
$E^+ [aa]$	$P_E (1-p)^2$	$R_E$	$P_E (1-p)^2 R_E / \Sigma$	$R_E$	$P_E (1-p)^2 R_E / \Sigma$
$E^- [AA]$	$(1-P_E) p^2$	$R_G$	$(1-P_E) p^2 R_G / \Sigma$	$R_G$	$(1-P_E) p^2 R_G / \Sigma$
$E^- [Aa]$	$(1-P_E) 2p(1-p)$	$R_G$	$(1-P_E) 2p(1-p) R_G / \Sigma$	1	$(1-P_E) 2p(1-p) / \Sigma$
$E^- [aa]$	$(1-P_E) (1-p)^2$	1	$(1-P_E) (1-p)^2 / \Sigma$	1	$(1-P_E) (1-p)^2 / \Sigma$

<sup>a</sup> Under dominant model:  $\Sigma = P_E (p^2 + 2p(1-p) R_E R_G R_I + P_E (1-p)^2 R_E + (1-P_E) (p^2 + 2p(1-p) R_G + (1-P_E) (1-p)^2)$ .

Under recessive model:  $\Sigma = P_E p^2 R_E R_G R_I + P_E (2p(1-p) + (1-p)^2) R_E + (1-P_E) (p^2 R_G + (1-P_E) ((1-p)^2 + 2p(1-p)))$ .

Table 2

Exposure and conditional genotype distribution of unaffected siblings according to case genotype for a dominant model. For a recessive model, for  $Aa$  siblings,  $(1-b)$  becomes  $(1-d)$  and  $(1-a)$  becomes  $(1-c)$  for each case genotype; otherwise, the equations are identical to those for the dominant model

Unaffected sib	Case genotype		
	[aa]	[Aa]	[AA]
$E^- [aa]$	$\left(p\left(\frac{p}{4}-1\right)+1\right)(1-P_E)(1-d)$	$\left(\frac{p(p-3)+2}{4}\right)(1-P_E)(1-d)$	$\left(\frac{(1-p)^2}{4}\right)(1-P_E)(1-d)$
$E^+ [aa]$	$\left(p\left(\frac{p}{4}-1\right)+1\right)P_E(1-c)$	$\left(\frac{p(p-3)+2}{4}\right)P_E(1-c)$	$\left(\frac{(1-p)^2}{4}\right)P_E(1-c)$
$E^- [Aa]$	$\left(\frac{p(2-p)}{2}\right)(1-P_E)(1-b)$	$\left(\frac{p(1-p)+1}{2}\right)(1-P_E)(1-b)$	$\left(\frac{(1-p)^2}{2}\right)(1-P_E)(1-b)$
$E^+ [Aa]$	$\left(\frac{p(2-p)}{2}\right)P_E(1-a)$	$\left(\frac{p(1-p)+1}{2}\right)P_E(1-a)$	$\left(\frac{(1-p)^2}{2}\right)P_E(1-a)$
$E^- [AA]$	$\left(\frac{p^2}{4}\right)(1-P_E)(1-b)$	$\left(\frac{p(p+1)}{4}\right)(1-P_E)(1-b)$	$\left(\frac{(1+p)^2}{4}\right)(1-P_E)(1-b)$
$E^+ [AA]$	$\left(\frac{p^2}{4}\right)P_E(1-a)$	$\left(\frac{p(p+1)}{4}\right)P_E(1-a)$	$\left(\frac{(1+p)^2}{4}\right)P_E(1-a)$

With:  $d = 0.001$ ;  $c = \frac{R_{Ed}}{1 + R_{Ed} - d}$ ;  $b = \frac{R_G d}{1 + R_G d - d}$ ;  $a = \frac{Rc(1-d)b}{d(1-b)(1-c) + R_G cb(1-d)}$ .

When there is a correlation in  $E$  between siblings,  $P_E$  becomes  $w$  if the case is not exposed to  $E$ , and  $m$  if the case is exposed to  $E$ .

controls, in the two designs. Both designs give unbiased estimates of  $R_I$ ; indeed, the coverage of the confidence intervals is at the nominal 95% level for all simulated scenarios (data not shown). Thus, when  $RE > 1$ , the case–combined–control design is more powerful than the corresponding classical case–control design; when  $RE < 1$ , the case–combined–control design is less powerful.

We studied the efficiency of the case–combined–control study relative to a classical case–control study according to different frequencies of  $G$  and  $E$  ( $P_G, P_E$ ), the main effect of  $G$  and  $E$  ( $R_G, R_E$ ), and the  $G \times E$  interaction effect ( $R_I$ ). We also examined different  $F$  (ranging from 0 to 2 to represent the average number of available sibling controls per case). Finally, we assessed efficiency when there was a correlation in  $E$  between siblings. Initial evaluations examined both dominant and recessive genetic models (see frequencies and main effects of  $G$  and  $E$ ). The results showed similar patterns for the two genetic models so subsequent evaluations (e.g., varying  $F$  and the correlation in  $E$  between siblings) were restricted to the dominant model.

## 4. Results

We present estimates of relative efficiency ( $RE$ ) in detecting  $G \times E$  interaction according to the parameters listed above.

### 4.1. According to the frequencies of $G$ and $E$

Fig. 1 presents  $RE$  for different frequencies of  $G$  for a dominant and recessive gene with  $R_G = 3$ ,  $R_E = 2$ ,  $R_I = 5$ ,  $P(E) = 0.2$ , and  $F = 0.5$ . The results show a rapid decrease in  $RE$  as  $P(G)$  increases up to 0.2, after which the decrease is slower and  $RE$  approaches 1 (0.99) when  $P(G) = 0.5$  (data not shown). For a dominant model,  $RE$  decreases from 1.26

when  $P(G) = 0.001$  to 1.08 when  $P(G) = 0.2$ . For a recessive model,  $RE$  is slightly lower than with a dominant model, decreasing from 1.18 when  $P(G) = 0.001$  to 1.07 when  $P(G) = 0.2$ . In contrast to the dramatic effect of  $P(G)$ , for most frequencies of  $E$ , there is little change in  $RE$  as  $P(E)$  changes, regardless of the values of the other parameters in the model (data not shown).

The gain and/or change in  $RE$  is insignificant for common genes and moderate main effects. Indeed, when  $R_I = R_G = R_E = 1.5$ ,  $RE = 1.05$  when  $P(G) = 0.01$ , and  $RE = 0.99$  when  $P(G) = 0.5$ .

### 4.2. According to the main effects of $E$ and $G$ and the interaction effect

Table 3 presents the relative efficiencies for different values of  $R_G$  (3, 10 or 1.5, 5) and  $R_E$  (1.5, 5) for a rare [ $P(G) =$

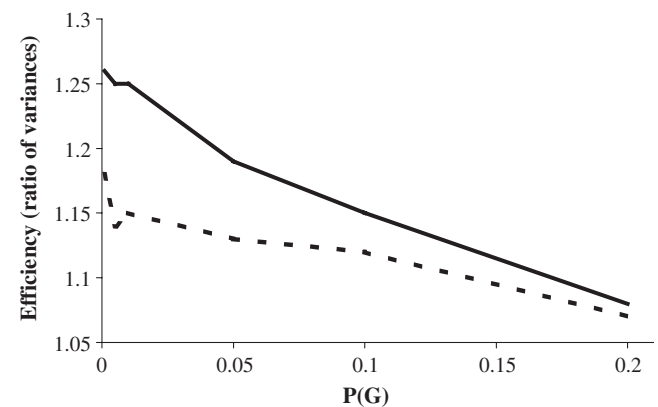


Fig. 1. Relative efficiency ( $RE$ ) according to the frequency of  $G$  for a dominant (bold line) and recessive gene (dashed line) with  $R_I = 5$ ,  $R_E = 2$ ,  $R_G = 3$ ,  $P(E) = 0.2$  and  $F = 0.5$ . \* $RE$  is defined as the ratio of the variance of  $\beta_I$  of the classical case–control study design divided by the variance of  $\beta_I$  of the case–combined–control design.

Table 3

Relative efficiency for  $G \times E$  interaction detection according to main and interaction effects

For a rare gene; $P(G) = 0.01$ ; $P(E) = 0.2$					For a common gene; $P(G) = 0.2$ ; $P(E) = 0.2$				
$R_E$	$R_G = 3$		$R_G = 10$		$R_E$	$R_G = 1.5$		$R_G = 5$	
	$R_I = 1.5$	$R_I = 5$	$R_I = 1.5$	$R_I = 5$		$R_I = 1.5$	$R_I = 5$	$R_I = 1.5$	$R_I = 5$
	Dominant gene					Dominant gene			
1.5	1.20	1.27	1.88	2.05	1.5	1.03	1.04	1.11	1.11
5	1.14	1.14	1.64	1.64	5	1.01	1.01	1.08	1.07
	Recessive gene					Recessive gene			
1.5	1.12	1.16	1.52	1.62	1.5	1.02	1.03	1.09	1.10
5	1.08	1.08	1.38	1.36	5	1.01	1.01	1.07	1.06

0.01] and a common [ $P(G) = 0.2$ ] dominant and recessive gene with  $P(E) = 0.2$  and  $R_I = 1.5$  or 5. For a rare gene, the results show an increase in  $RE$  as  $R_G$  increases. The relative efficiencies in general also slightly increase when  $R_I$  increases. As was previously described, when the genetic model is recessive,  $RE$  show the same trend as for the dominant model, but with reduced magnitudes for a given  $P(G)$ . For a common dominant or recessive gene, there is essentially no effect of  $R_G$ ,  $R_E$ , or  $R_I$  on  $RE$  and there is also very little gain in power compared with the classical case-control design.

Table 3 also shows that for a given  $R_G$ ,  $RE$  decreases as  $R_E$  increases. For example, for a rare dominant gene when  $R_G = 3$ ,  $RE$  decreases from 1.27 when  $R_E = 1.5$  to 1.14 when  $R_E = 5$  (and 1.05 when  $R_E = 10$ ) for  $R_I = 5$ .

#### 4.3. According to the number of available sibling controls per case

Fig. 2 shows the effect of the number of available sibling controls per case on  $RE$  for a rare [ $P(G) = 0.01$ ] and a common [ $P(G) = 0.2$ ] dominant gene. For this evaluation,  $F$  varies from 0–2 sib controls per case. All other parameters are fixed with  $R_G = 3$ ,  $R_E = 2$ ,  $R_I = 5$ ,  $P(E) = 0.2$ . A 1:(1+ $F$ ) case-combined-control design is always more powerful than a 1:(1+ $F$ ) classical case-unrelated-control

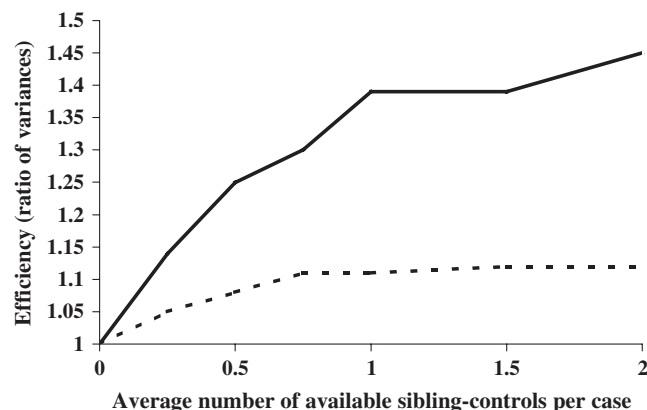


Fig. 2. Relative efficiency ( $RE$ ) according to the average number of available sibling controls per case for a rare [ $P(G) = 0.01$ ; bold line] and common [ $P(G) = 0.2$ ; dashed line] dominant gene with  $P(E) = 0.2$ ,  $R_G = 3$ ,  $R_E = 2$ ,  $R_I = 5$ .

design.  $RE$  increases as the number of available siblings per case increases. For example, when  $P(G) = 0.01$ ,  $RE$  increases from 1.14 when 25% of cases have one sibling control to 1.40 when 100% of cases have one sibling control and to 1.45 when 100% of cases have two sibling controls. In other words, a classical 1:1.5 case-control design requires 1.25 times more cases than a 1:1.5 case-combined-control design. However, there is little gain in power compared with the classical case-control design for a common gene [ $P(G) = 0.2$ ].

#### 4.4. According to the correlation between sibs' exposure ( $E$ )

Fig. 3 examines the  $RE$  according to different values of  $P(E)$  with  $R_G = 3$ ,  $R_E = 1.5$ ,  $R_I = 5$ , and  $P(G) = 0.01$  for four different correlations of  $E$  between sibs' exposure ( $OR_{EC} = 1, 2, 3, 5$ ). When there is no correlation in  $E$  exposure between sibs ( $OR_{EC} = 1$ ),  $RE$  slightly decreases as  $P(E)$  increases, but there is little overall change in  $RE$  [ $RE$  of 1.32 at  $P(E) = 0.05$  vs. 1.26 at  $P(E) = 0.5$ ]. In contrast, when  $OR_{EC} > 1$ , that is, when there is a correlation in  $E$  between sibs' exposures, there is a more pronounced decrease in  $RE$  as  $P(E)$  increases. For higher values of  $OR_{EC}$ ,  $RE < 1$  when  $P(E) \geq 0.5$  (data not shown). In addition, as  $OR_{EC}$  increases,  $RE$  decreases slightly from 1.32 when

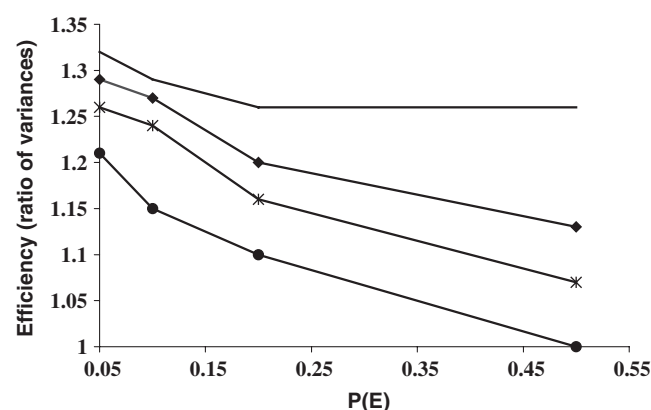


Fig. 3. Relative efficiency ( $RE$ ) according to different values of  $P(E)$  for four different correlations of  $E$  between sibs' exposure ( $OR_{EC} = 1$ , bold line;  $OR_{EC} = 2$ , diamond line;  $OR_{EC} = 3$ , starred line;  $OR_{EC} = 5$ , circled line) with  $R_E = 1.5$ ,  $R_G = 3$ ,  $R_I = 5$ , and  $P(G) = 0.01$ .

$OR_{EC} = 1$  to 1.21 when  $OR_{EC} = 5$  for  $P(E) = 0.05$ .  $RE$  decreases more sharply for higher frequencies of  $P(E)$ . For moderate main and interaction effects, for example,  $R_E = R_G = R_I = 1.5$ , there is again very little gain in  $RE$ . In addition, there is marginal change in  $RE$  as  $OR_{EC}$  or  $P(E)$  increases. Finally, strong  $E$  correlation can induce relative efficiencies less than 1 [e.g.,  $OR_{EC} = 5$  and  $P(E) \geq 0.2$ ] (data not shown).

#### 4.5. Feasibility of the case–combined–control design

To evaluate the feasibility of the case–combined–control design in  $G \times E$  interaction assessment, power for different sample sizes are calculated for different values of  $R_I$ ,  $R_G$ ,  $R_E$ , with  $F = 0.5$ ,  $P(E) = 0.2$ , and  $OR_{EC} = 1$  for a rare [ $P(G) = 0.01$ ] and a common [ $P(G) = 0.2$ ] dominant gene. Power is calculated using a two-sided test at the 5% level for type I error. Table 4 shows power calculations for a classical 1:1.5 case–control design (named traditional) and a 1:1.5 case–combined–control design (named combined). When  $G$  is rare and  $G \times E$  interaction is moderate (e.g.,  $R_I = 1.5$ ), the required sample size is very large ( $>20,000$  cases and 30,000 controls) reaching unrealistic numbers. Although the case–combined–control design is more efficient than the traditional design, the needed sample size is only realistic when there are strong  $G \times E$  interaction (e.g.,  $R_I \geq 3$ ) and  $G$  (e.g.,  $R_G \geq 3$ ) effects. For example, if one were interested in studying the interaction between mutations in the Ataxia-Telangiectasia (ATM) gene (estimated prevalence for truncating mutations of about 1% in the general population and a three- to fourfold increased risk of

breast cancer) [28] and ionizing radiation of the chest ( $R \sim 1.5$ ), a traditional case–control study would require 3,180 breast cancer cases and 4,770 controls (i.e., a total of 7,950 women) to detect an interaction of 3 with 81% power. In contrast, the case–combined–control design would require 2,500 breast cancer cases, 2,500 unrelated (population/hospital) controls and about 1,250 unaffected sisters (i.e., a total of 6,250 women); that is, 1,700 (about 21%) subjects less than for the traditional case–control study. When  $R_I = 5$  and power = 80%, the sample size decreases to 1,360 breast cancer cases and 2,040 controls (i.e., 3,400 women in total) in the traditional case–control study and 1,020 breast cancer cases, 1,020 unrelated controls, and about 510 unaffected sisters (i.e., 2,550 total women), 850 women or about 25% fewer subjects than in the traditional design. To further illustrate these examples, Table 5 shows approximations of the expected cell counts for the joint distribution of the ionizing radiation exposure ( $E$ ) and the ATM genotype ( $G$ ) in cases, unrelated controls, and sibling controls. The proportion of cases and controls (unrelated or sibling) jointly exposed to ionizing radiation and having a mutation (i.e.,  $E^+G^+$ ) is small (e.g., 2.4, 0.2, and 0.5%, respectively, when  $R_I = 3$ ). When  $G$  is common [e.g.,  $P(G) = 0.2$ ], the required sample size remains realistic ( $\leq 2,200$  cases and 3,300 controls). In general, though, the gain in power for the combined design relative to the traditional design is less for the common gene than for the rare gene. Indeed, to illustrate, if one were now interested in steroid hormone metabolism genes (prevalence of polymorphisms of about 0.2 and  $R_G = 1.5$ ) and their potential interactions with reproductive factors in breast

Table 4

Power for  $G \times E$  interaction detection (two-sided test at the 5% level) according to different main and interaction effects for a traditional case–control study [traditional] and a case–combined–control study [combined]

A rare dominant gene; $P(G) = 0.01$ ; $P(E) = 0.2$						A common dominant gene; $P(G) = 0.2$ ; $P(E) = 0.2$					
$R_G = 3$ ; $R_E = 1.5$ ; $R_I = 5$						$R_G = 3$ ; $R_E = 1.5$ ; $R_I = 3$					
No. of cases	500	750	1,000	1,250	1,500	No. of cases	150	250	350	450	
Traditional	22	43	62	77	84	Traditional	46	70	83	92	
Combined	33	61	79	90	94	Combined	52	76	87	95	
$R_G = 3$ ; $R_E = 1.5$ ; $R_I = 3$						$R_G = 1.5$ ; $R_E = 1.5$ ; $R_I = 3$					
No. cases	1,700	2,100	2,500	2,900	3,300	No. of cases	200	300	400	500	
Traditional	50	61	70	77	83	Traditional	61	78	90	95	
Combined	63	73	81	88	92	Combined	62	83	92	95	
$R_G = 1.5$ ; $R_E = 1.5$ ; $R_I = 3$						$R_G = 3$ ; $R_E = 1.5$ ; $R_I = 1.5$					
No. cases	2,000	2,500	3,000	3,500	4,000	No. cases	1,600	1,800	2,000	2,200	2,400
Traditional	51	62	71	78	83	Traditional	64	70	73	78	81
Combined	55	66	75	83	86	Combined	70	75	75	83	85
$R_G = 3$ ; $R_E = 1.5$ ; $R_I = 1.5$						$R_G = 1.5$ ; $R_E = 1.5$ ; $R_I = 1.5$					
No. of cases	15,000	18,000	21,000	24,000	27,000	No. of cases	1,400	1,600	1,800	2,000	2,200
Traditional	57	63	71	73	79	Traditional	58	64	69	74	78
Combined	66	72	82	86	87	Combined	59	65	71	76	80
$R_G = 1.5$ ; $R_E = 1.5$ ; $R_I = 1.5$											
No. of cases	20,000	25,000	30,000	35,000	40,000						
Traditional	60	68	76	84	86						
Combined	61	73	77	85	89						

Percent power shown for the number of cases listed under each traditional and combined model comparison.

Table 5  
Illustrative examples from breast cancer: expected cell counts for the joint distribution of the exposure ( $E$ ) and the genotype ( $G$ ) in cases, controls (unrelated and sibling) in each design

Design	Cases						Siblings						Cases/controls	
	$E^- G^-$	$E^- G^+$	$E^+ G^-$	$E^+ G^+$	$E^- G^-$	$E^- G^+$	$E^+ G^-$	$E^+ G^+$	$E^- G^-$	$E^- G^+$	$E^+ G^-$	$E^+ G^+$	Total	Total
Rare gene: $P(G) = 0.01$ ; $P(E) = 0.2$ ; $R_G = 3$ ; $R_E = 1.5$ ; $R_I = 3$ ( $1 - \beta$ ) = 0.81														
Traditional	2210 (69.5%)	67 (2.1%)	830 (26.1%)	73 (2.3%)	3,778 (79.2%)	38 (0.8%)	944 (19.8%)	10 (0.2%)	—	—	—	—	3,180/4,770	3,180/4,770
Combined	1,738 (69.5%)	52 (2.1%)	652 (26.1%)	58 (2.3%)	1,980 (79.2%)	20 (0.8%)	495 (19.8%)	5 (0.2%)	973 (77.8%)	27 (2.2%)	243 (19.5%)	7 (0.5%)	2,500/3,750	2,500/3,750
Rare gene: $P(G) = 0.01$ ; $P(E) = 0.2$ ; $R_G = 3$ ; $R_E = 1.5$ ; $R_I = 5$ ( $1 - \beta$ ) = 0.80														
Traditional	930 (68.4%)	29 (2.1%)	348 (25.6%)	53 (3.9%)	1,616 (79.2%)	16 (0.8%)	404 (19.8%)	4 (0.2%)	—	—	—	—	1,360/2,040	1,360/2,040
Combined	698 (68.4%)	21 (2.1%)	261 (25.6%)	40 (3.9%)	808 (79.2%)	8 (0.8%)	202 (19.8%)	2 (0.2%)	394 (77.2%)	14 (2.8%)	98 (19.3%)	4 (0.7%)	1,020/1,530	1,020/1,530
Common gene: $P(G) = 0.2$ ; $P(E) = 0.2$ ; $R_G = 1.5$ ; $R_E = 1.5$ ; $R_I = 3$ ( $1 - \beta$ ) = 0.83														
Traditional	157 (45.9%)	59 (17.3%)	59 (17.3%)	67 (19.5%)	328 (63.9%)	83 (16.1%)	82 (16.0%)	20 (4.0%)	—	—	—	—	342/513	342/513
Combined	138 (45.9%)	52 (17.3%)	52 (17.3%)	58 (19.5%)	192 (63.9%)	48 (16.0%)	48 (16.0%)	12 (4.0%)	86 (57.4%)	34 (22.6%)	22 (14.4%)	8 (5.6%)	300/450	300/450
Common gene: $P(G) = 0.2$ ; $P(E) = 0.2$ ; $R_G = 1.5$ ; $R_E = 1.5$ ; $R_I = 1.5$ ( $1 - \beta$ ) = 0.80														
Traditional	1,171 (50.9%)	442 (19.2%)	439 (19.1%)	248 (10.8%)	2,205 (63.9%)	555 (16.1%)	552 (16.0%)	138 (4.0%)	—	—	—	—	2,300/3,450	2,300/3,450
Combined	1,120 (50.9%)	422 (19.2%)	420 (19.1%)	238 (10.8%)	1,406 (63.9%)	354 (16.1%)	352 (16.0%)	88 (4.0%)	660 (60.0%)	220 (20.0%)	165 (15.0%)	55 (5.0%)	2,200/3,300	2,200/3,300

cancer ( $R_E \sim 1.5$ ) [29], then when  $R_I = 3$ , power will be 83% with 342 breast cancer cases and 513 controls in the traditional design, and 300 breast cancer cases, 300 unrelated controls, and about 150 unaffected sisters in the combined design. In this example, the combined design required 12% fewer women than the traditional design. The joint distribution of a reproductive factor ( $E$ ) and the steroid hormone metabolism genotype ( $G$ ) in cases, unrelated controls, and sibling controls is further illustrated in Table 5 for both  $R_I = 1.5$  and  $R_I = 3$ .

## 5. Discussion

The case–combined–control design using both population-based and related-controls for studies to detect  $G \times E$  interaction is more efficient than a classical case–control study for interaction detection involving less common events (i.e., frequencies  $\leq 0.2$ ). The parameters that are the most important for determining relative efficiency are the number of available sibling controls per case and the frequencies of the genetic factor of interest. Relative efficiencies decrease as the frequency of  $G$  increases. Dominant and recessive genetic models show similar patterns but with smaller  $RE$  gains for the recessive model. Also, as a correlation in  $E$  exposure between sibs increases,  $RE$  decreases. This decrease is greater for higher frequencies of  $E$ . Finally, for common genes and moderate main and interaction effects (e.g.,  $R_E = R_G = R_I = 1.5$ ) with or without a correlation in  $E$  between siblings, the case–combined–control design is often less efficient than a classical case–control study.

However, to fully evaluate a study design, efficiency needs to be complemented by feasibility as measured by the required sample size. When  $G$  is rare [i.e.,  $P(G) = 0.01$ ] and the interaction value is small (i.e.,  $< 3$ ), the case–combined–control design is more efficient than a classical case–control design but the corresponding required sample size is very large and remains prohibitive (i.e.,  $> 20,000$  cases and 30,000 controls). For smaller frequencies of  $G$ , as might be observed for major genes in cancer or other chronic diseases, the needed sample size might be realistic only when the environmental factor is common and the interaction effect is high (i.e.,  $\geq 5$ ). For more common genetic factors [i.e.,  $P(G) = 0.2$ ], the required sample size is realistic but the gain in relative efficiency of the case–combined–control design may be minimal. The range of scenarios where the case–combined–control design is both relatively efficient and feasible may thus appear narrow. However, this interval includes numerous polymorphic genes and inside this range, this design appears to be a useful alternative to study  $G \times E$  interaction where some classical approaches remain unrealistic.

Previous investigators have argued against using multiple control groups. If one control group is superior to all practical alternatives, then efforts should not be wasted on collecting controls from multiple sources with multiple required infrastructures [23]. However, there are situations when using

multiple control groups might be warranted. In this article, we have presented a modification to the case–control design that ascertains two different types of controls—related and unrelated controls. The purpose of selecting two control groups is to maximize the efficiency and feasibility for examining  $G \times E$  interactions. This design requires several assumptions including homogeneity between the two types of controls with respect to the variables involved in the interactions, no population stratification bias, no difference in the distribution of variables of interest between cases with and without sibling controls, and exchangeability of the case–sibling covariates under investigation.

For purposes of presentation and ease of comparison, conditional logistic regression analyses were performed on simulated data that included a correlated genetic and environmental factor in cases and sibling controls. Under this scenario, the estimates were unbiased. However, in real situations with greater complexity, bias may result if important correlations are ignored. In such situations, other analytic approaches, such as polytomous regression, might be required. In addition, incorporating weighting into the analyses to account for the differences in the ascertainment of the two control groups, may require fewer assumptions than conditional logistic regression and modify the precision. Examination of the reliability and validity of conditional logistic regression when there is deviation from the required assumptions and evaluation of different analytical approaches is planned.

The critical assumptions for this design are not testable before the data has been collected. If one or more assumptions were not valid, the proposed analysis for examining  $G \times E$  interaction would not be appropriate. For example, if there is evidence for heterogeneity in the interaction odds ratios from the two types of controls, then it is not appropriate to conduct a combined analysis and alternative analytic strategies such as using polytomous regression or stratification or performing separate analyses will be required. In addition, if the environmental covariates in cases and sibling controls are not exchangeable or there are differences in the distribution of variables of interest between cases with and without sibling controls, then the case–sibling control analysis will be subject to bias [24,25]. The extent of this bias is currently unknown. These issues will be further investigated.

There has been much discussion in the literature about possible biases from using unrelated controls to examine gene–disease associations and/or gene–environment interactions (for review, see [30,31]). Although this potential population stratification bias could produce problems in the case–combined–control design, recent studies have shown that, in general, population stratification produces a minimal and tolerable bias [30,32]. For the rare study in which population stratification bias is a major concern, minimization of potential bias may be accomplished by controlling for ethnicity [33] or by using approaches such as genomic control [34] or modeling population substructure [35,36].

Other study designs and analytical strategies have been proposed to examine interaction involving rare factors (e.g., [7–10,13,37–40]). For approaches that permit estimation of both main and interaction effects, the principle of these designs is similar to the case–combined–control design, that is, increasing the frequency of the rare factors through oversampling, to increase the power of the study. One group of designs used unrelated controls. Flexible matching strategies [10] with varying proportions of an environmental matching factor among selected controls increased the power and efficiency to detect  $G \times E$  interactions in case–control studies. The highest efficiency was observed for a rare exposure that was a strong risk factor. However, this design is not recommended if the main effect of the matching factor has not been thoroughly studied or if one is interested in additive risk interactions. Sturmer and Brenner [9,10] were unable to fully evaluate rare factors with prevalences  $<0.05$ ; thus, the efficiency, and more importantly, the feasibility of this design for these scenarios is not available.

Multistage designs, including “countermatched designs” [7,41] and “balanced designs” [8,39,42,43] also have been proposed to study rare factors. These designs employ methods of sampling from an at-risk population for nested case–control studies. In the balanced design, one selects both cases and controls to oversample the rare factor of interest. The oversampling is taken into account in the analysis to obtain unbiased estimates of the effects. The efficiency for estimating exposure covariate (e.g.,  $G \times E$ ) interaction has not been fully investigated for the balanced design. More extensive efficiency evaluation has been conducted using the countermatched design [11]. Limited direct comparisons between the balanced and countermatching designs, however, showed similar efficiencies in interaction estimation [41,43].

The second group of designs used related subjects as controls. Various relative designs have been proposed for examining main effect(s) and  $G \times E$  interactions including case–parent (e.g., [14,37,38,44]), case–cousin–control [12,13,37,40], and case–sibling–control [12–14,37,40]. Few such designs have been evaluated for  $G \times E$  interaction assessment [12–14,40]. In general, relative control subjects were less efficient than population-based control subjects for detecting the genetic factor main effect, except when cases with a positive family history were oversampled [40]. However, relative control subjects were the most efficient group for detecting interaction. The gain in efficiency, however, decreased as the frequency of  $G$  increased as we also observed in the case–combined–control design. A major difference between these prior studies and the current article is the *a priori* defined number of siblings per case. These prior studies assumed 1:1 or 1:2 case:sibling–control ratios. Limited examinations have suggested that approximately 50% of cases may have appropriate sibling controls ([20,45], unpublished data). Thus, the 1:1 or 1:2 matching may be unrealistic in many situations or may lead to the exclusion of numerous cases.

Theoretically, the case–combined–control design would be useful for examining interaction. In practice, however, we may find that there are few situations where combining control groups for all variables of interest are possible. Future studies will examine whether the case–combined–control design would still offer advantages over a single case–unrelated–control or case–related–control study design if combining the control groups were not appropriate. Although the case–combined–control design has increased complexity for control recruitment, it has the potential to exploit the collection of two different types of controls, related and unrelated. Using siblings as controls substantially increases the power for some schemas of  $G \times E$  interaction. Adding unrelated controls counterbalances the loss of efficiency from a design that recruits only sibling controls (because of the availability of only a fraction of sibling controls and the major assumptions required for these sibling controls).

### Acknowledgments

This project was supported by INSERM, the U.S. National Cancer Institute, the Foundation Philippe Inc., the Association pour la Recherche contre le Cancer and the University of Paris Sud. This work was conducted while Nadine Andrieu was a guest researcher at the Genetic Epidemiology Branch of the National Cancer Institute. The authors thank Marie-Gabrielle Dondon, Inserm, Institut Curie and Angela Fahey, IMS, for their assistance with simulations and computations.

### Appendix A

Details of calculations are given below.

#### Probability that the case's sibling has $G$

Table 2 shows the conditional probability that a case's sibling has  $G$ , given the case has  $G$  [i.e.,  $P(G_S = G + | G_C = G + )$ ] for the three case genotypes. For a dominant model, exposure to  $G$  occurs when a subject is  $AA$  or  $Aa$ . For a recessive model, exposure to  $G$  is equivalent to  $AA$  only. To determine the probability for each sibling control, we incorporate  $P_E$  and the risk of disease given the  $E$  and  $G$  status of each control. For example, under a dominant model with case  $[aa]$ ,  $P(\text{Sib} = E^-, aa, D^- | \text{Case} = aa) =$

$$\left( p \left( \frac{p}{4} - 1 \right) + 1 \right) (1 - P_E) (1 - d)$$

where  $d = 0.001$ , baseline disease risk.

A similar approach was used for the recessive model.

#### Probability that the cases' sibling is exposed to $E$

We define  $m$  as the probability that a case's sibling is exposed to  $E$  if the case is exposed to  $E$  [ $P(E_S = E^+ | E_C = E^+)$ ]. Similarly, we define  $(1 - w)$  as the probability that a sibling has not been exposed to  $E$  given the case has not been exposed [ $P(E_S = E^- | E_C = E^-)$ ] [23].

Given exchangeability for  $E$ , the frequency of  $E$  is the same in the case and her sibling control, thus uniquely determining the joint exposure distribution between the two siblings by constraining the marginal probabilities to be equal. Thus,

$$w = \frac{P_E(1 - m)}{1 - P_E}.$$

We use the following equation to define the exposure relationship between a case and his/her sibling control,

$$m = \frac{OR_{EC} P_E}{1 - P_E + OR_{EC} P_E}.$$

When  $OR_{EC} = 1$ ,  $m = P_E$ , and there is no correlation in  $E$  between siblings. We examined the following scenarios,  $OR_{EC} = 1, 2, 3$ , and  $5$ . We incorporated  $OR_{EC}$  into the siblings' probabilities (see probabilities that the case's sibling has  $G$  above) by adding  $m$ ,  $1 - m$ ,  $w$ , or  $1 - w$  to the equation, as appropriate, depending on the  $E$  status of the case and matched sibling. For example, if  $E_C = E^-$  and  $E_S = E^+$ , under a dominant model with case  $[aa]$ ,

$$\begin{aligned} P(\text{Sib} = E^+, aa, D^- | \text{Case} = E^-, aa) \\ = \left( p \left( \frac{p}{4} - 1 \right) + 1 \right) w (1 - d). \end{aligned}$$

Similarly,

$$\begin{aligned} P(\text{Sib} = E^+, AA, D^- | \text{Case} = E^+, AA) \\ = \left( \frac{(1 + p)^2}{4} \right) m (1 - a). \end{aligned}$$

### References

- [1] Shilberg O, Dorman JS, Ferrell RE, Trucco M, Shahar A, Kuller LH. The next stage: molecular epidemiology. *J Clin Epidemiol* 1997;50: 633–638.
- [2] Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene–environment interaction: case–control studies with no controls! *Am J Epidemiol* 1996;144:207–13.
- [3] Yang Q, Khoury MJ. Evolving methods in genetic epidemiology, III. Gene–environment interaction in epidemiologic research. *Epidemiol Rev* 1997;19:33–43.
- [4] Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene–environment interaction: an overview of available methods. *Epidemiol Rev* 1998;20:137–47.
- [5] Smith PG, Day NE. The design of case–control studies: the influence of confounding and interaction effects. *Int J Epidemiol* 1984;13:356–65.
- [6] Attia J, Thakkinian A, D'Este C. Meta-analyses of molecular association studies: methodologic lessons for genetic epidemiology. *J Clin Epidemiol* 2003;56:297–303.
- [7] Langholz B, Clayton D. Sampling strategies in nested case–control studies. *Environ Health Perspect* 1994;102(suppl 8):47–51.
- [8] Breslow NE. Case–control study, two phase. New York: Wiley; 1998.
- [9] Sturmer T, Brenner H. Potential gain in efficiency and power to detect gene–environment interactions by matching in case–control studies. *Genet Epidemiol* 2000;18:63–80.
- [10] Sturmer T, Brenner H. Flexible matching strategies to increase power and efficiency to detect and estimate gene–environment interactions in case–control studies. *Am J Epidemiol* 2002;155:593–602.

- [11] Andrieu N, Goldstein A, Thomas D, Langholz B. Counter-matching in studies of gene–environment interaction: efficiency and feasibility. *Am J Epidemiol* 2001;153:265–74.
- [12] Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. *J Natl Cancer Inst Monogr* 1999;31–7.
- [13] Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case–control studies of candidate genes and gene–environment interactions: basic family designs. *Am J Epidemiol* 1999;149:693–705.
- [14] Gauderman WJ. Sample size requirements for matched case–control studies of gene–environment interaction. *Stat Med* 2002;21(1):35–50.
- [15] Gladen BC. Matched-pair case–control studies when risk factors are correlated within the pairs. *Int J Epidemiol* 1996;25:420–5.
- [16] Andrieu N, Goldstein AM. Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. *Int J Epidemiol* 1996;25:649–57.
- [17] Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case–control studies. II. Types of controls. *Am J Epidemiol* 1992;135:1029–41.
- [18] Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case–control studies. III. Design options. *Am J Epidemiol* 1992;135:1042–50.
- [19] Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case–control studies, I. Principles. *Am J Epidemiol* 1992;135:1019–28.
- [20] Andrieu N, Demenais F. Interactions between genetic and reproductive factors in breast cancer risk in a French family sample. *Am J Hum Genet* 1997;61:678–90.
- [21] Botto LD, Khoury MJ. Commentary: facing the challenge of gene–environment interaction: the two-by-four table and beyond. *Am J Epidemiol* 2001;153:1016–20.
- [22] Austin H, Flanders WD, Rothman KJ. Bias arising in case–control studies from selection of controls from overlapping groups. *Int J Epidemiol* 1989;18:713–6.
- [23] Rothman KJ, Greenland S. *Modern epidemiology*. 2nd ed. Philadelphia: Lippincott-Raven; 1998.
- [24] Langholz B, Ziogas A, Thomas DC, Faucett C, Huberman M, Goldstein L. Ascertainment bias in rate ratio estimation from case–sibling control studies of variable age-at-onset diseases. *Biometrics* 1999;55:1129–36.
- [25] Goldstein AM, Hodge SE, Haile RW. Selection bias in case–control studies using relatives as the controls. *Int J Epidemiol* 1989;18:985–9.
- [26] Stata Corp. *Stata statistical software: release 7.0*. College Station, TX: Stata Corporation; 2001.
- [27] Breslow NE, Day NE. *The analysis of case–control studies*. In: Davis W, editor. *Statistical methods in cancer research*. Lyon, France: International Agency for Research on Cancer; 1980.
- [28] Geoffroy-Perez B, Janin N, Ossian K, Laugé A, Stoppa-Lyonnet D, Andrieu N. Variation in breast cancer risk of heterozygotes for Ataxia-Telangiectasia according to environmental factors. *Int J Cancer* 2002;99:619–23.
- [29] Dunning AM, Healey CS, Pharoah PD, Teare MD, Ponder BA, Easton DF. A systematic review of genetic polymorphisms and breast cancer risk. *Cancer epidemiology. Biomarkers Prev* 1999;8:843–54.
- [30] Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000;92:1151–8.
- [31] Thomas DC. Design of gene characterization studies: an overview. *J Natl Cancer Inst Monogr* 1999;17–23.
- [32] Millikan RC. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2001;93:156–8.
- [33] Caporaso N, Rothman N, Wacholder S. Case–control studies of common alleles and environmental factors. *J Natl Cancer Inst Monogr* 1999;25–30.
- [34] Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004.
- [35] Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;65:220–8.
- [36] Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case–control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001;68:466–77.
- [37] Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* 2000;152(3):197–203.
- [38] Umbach DM, Weinberg CR. The use of case–parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 2000;66(1):251–61.
- [39] White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am J Epidemiol* 1982;115:119–28.
- [40] Siegmund KD, Langholz B. Stratified case sampling and the use of family controls. *Genet Epidemiol* 2001;20:316–27.
- [41] Langholz B, Borgan O. Counter-matching: a stratified nested case–control sampling method. *Biometrika* 1995;82:69–79.
- [42] Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. *Am J Epidemiol* 1988;128:1198–206.
- [43] Breslow NE, Cain KC. Logistic regression for two-stage case–control data. *Biometrika* 1988;75:11–20.
- [44] Schaid DJ. Case–parents design for gene–environment interaction. *Genet Epidemiol* 1999;16:261–73.
- [45] Becher H, Schmidt S, Chang-Claude J. Reproductive factors and familial predisposition for breast cancer by age 50 years. A case–control–family study for assessing main effects and possible gene–environment interaction. *Int J Epidemiol* 2003;32:38–48.